Вестник Тюменского государственного университета.
2014. № 7. Физико-математические науки. Информатика. 199-204

© С.В. ДРОНОВ, А.С. САЗОНОВА
Алтайский государственный университет (Барнаул)
dsv@math.asu.ru, antonina1282@mail.ru

УДК 519.237

*ОБРАТНАЯ POST-HOC ЗАДАЧА КЛАСТЕРНОГО АНАЛИЗА И ЕЕ ПРИМЕНЕНИЕ К ДИСКРИМИНАЦИИ ДАННЫХ**

INVERSE POST-HOC PROBLEM OF THE CLUSTER ANALYSIS WITH APPLICATION TO DATA DISCRIMINATION

АННОТАЦИЯ. Рассматривается задача определения информационной важности статистических показателей объектов некоторого множества. Показатели объектов допускаются к рассмотрению как числовые, так и качественные категорированные. Решается задача: используя априорную информацию о порядке следования кластеров, ранжировать показатели по степени их важности. Предполагая существование дискриминационной функции, правильно разделяющей объекты по имеющимся кластерам, разработан алгоритм, позволяющий для каждого показателя X определить вид такого его преобразования f_X, что после замены в дискриминационной функции X на f_X его влияние на кластерную структуру множества выделяется оптимальным образом.

SUMMARY. We consider the problem of informational importance of characteristics determination for a set of clusterized objects. These characteristics can be either numerical ones or non-numerical with categories. Using priori information about the natural order of clusters, we propose the way to range characteristics with respect to the degree of their importance. Assuming the discrimination function separating correctly the objects into the available clusters, we developed a new algorithm. The algorithm defines a type of the proper f_X transformation for each X characteristic. In this case, if we replace the discriminatory function X with f_X, then a new discrimination function will show the influence of the characteristic on cluster structure of the set of objects in the optimal way.

КЛЮЧЕВЫЕ СЛОВА. Кластеризация, дискриминационная функция, кластерная переменная.

KEY WORDS. Clusterization, discriminatory function, cluster variable.

* *Работа выполнена в рамках программы стратегического развития ФГБОУ ВПО «Алтайский государственный университет» на 2012-2016 годы «Развитие Алтайского государственного университета в целях модернизации экономики и социальной сферы Алтайского края и регионов Сибири» (мероприятие «Конкурс грантов-2014», № 2014.312.1.4)*

Вводные замечания. Кластерный анализ в наше время получил широкое применение во многих областях исследовательской деятельности, например, в области медицины, социологии, психологии и др. В задачах кластеризации иногда смысл получаемых кластеров заранее известен. В медицине, например, это могут быть различные стадии диагностируемого заболевания, как правило, легко располагаемые по нарастанию его тяжести, в задачах проверки качества — степень удаленности от некого идеального образца (сорт изделия) и т.д. Из приведенных примеров ясно, что часто мы можем установить порядок следования кластеров, а следовательно, предполагать, что заданы их числовые метки, по крайней мере, в ранговой шкале, возрастающие в естественном порядке.

Условимся считать, что исследуемые объекты заданы какими-то своими показателями, и кластеры построены именно в соответствии со значениями этих показателей, а не по каким-то иным соображениям. Естественным образом возникает задача ранжирования имеющихся показателей по отношению к заданным числовым кластерным меткам. Действительно, установив путем решения этой задачи порядок важности формирующих кластеры показателей, мы сумеем определить степень важности каждого из них в диагностике, например, тяжести заболевания. Поставим задачу строго.

Постановка задачи. Пусть задано n объектов, каждый из которых имеет p числовых показателей $X_1,...,X_p$, и q качественных (нечисловых) категорированных показателей Y_i с s_i категориями соответственно, $i = 1,...,q$. Предположим, имеется некоторое значимое с точки зрения практики разбиение рассматриваемых объектов на m кластеров. Мы не будем задаваться здесь вопросом о том, как именно построен каждый из кластеров, но нам известен его «объективный» (экспертный) ранг, который мы временно примем за числовую метку соответствующего кластера.

Обозначим через $N(j)$ множество номеров тех объектов, которые составляют j-й кластер, $j=1,..,m$. Подобно тому, как это было сделано в [1], определим для каждого из объектов значение кластерной переменной, а именно: поставим каждому из объектов в соответствие номер того кластера, в который он отнесен. Т.о., построено отображение f из набора номеров объектов $\{1,...,n\}$ на множество всех имеющихся кластеров, и тем самым каждому объекту придана новая числовая характеристика. Значение $f(j)$ этой характеристики для j-го объекта будем называть кластерной переменной.

Задача определения информационной важности показателей, а также ранжирования показателей в соответствии со степенью их важности в современной статистической литературе называется post-hoc задачей кластерного анализа.

Рассмотрим post-hoc задачу определения значимости показателей $X_1,...,X_p, Y_1,...,Y_q$ заданных объектов. Для решения такой задачи предлагается предварительно произвести оцифровку качественных показателей $Y_1,...,Y_q$, т.е. присвоить категориям качественных показателей цифровые метки, которые будут отражать истинные различия между категориями.

Понятие «истинные различия» здесь, конечно же, нуждается в уточнении. Как известно, естественный способ задания различий между качественными показателями — составление таблиц их сопряженности. Потребуем, чтобы задаваемые метки были бы согласованы с совместными частотами встречаемости каждого из сочетаний категорий признаков. Такие метки назовем частотно-согласованными, следуя терминологии, предложенной в [2].

Способ построения меток, согласованных с таблицами сопряженности, известен под названием *анализ соответствий*. Подробное изложение этого способа можно найти, например, в [4]. В результате работы анализа соответствий каждая из категорий показателей может получить векторную метку размерности до $s = \min\limits_{1 \le i \le m}\{s_i\} - 1$ включительно. Поскольку координаты векторных меток формируются в порядке степени их разброса, то мы, как и в [3], выберем в качестве числовых меток первые координаты получающихся векторных меток (как наиболее информативные). После выполнения этих действий у каждого из рассматриваемых объектов будет иметься $p+q$ числовых показателей $X_1,...,X_p$, $X_{p+1},...,X_{p+q}$. Здесь через $X_{p+1},...,X_{p+q}$ обозначены «числовые варианты» первоначально заданных категорированных качественных показателей $Y_1,...,Y_q$.

Одним из наиболее часто применяемых показателей взаимозависимости двух случайных величин является парный коэффициент корреляции (см., например, [5]). Попробуем воспользоваться этой характеристикой для решения нашей задачи. Вычислим коэффициенты корреляции ρ_j между показателем X_j и кластерной переменной f, $j = 1,..., p+q$. Составим убывающий ряд из модулей найденных коэффициентов корреляции. Будем ранжировать значимость показателей по убыванию $|\rho_j|$, т.е. будем считать, что чем раньше в данном ряду встречается коэффициент, соответствующий какому-либо показателю, тем более важную роль в построении кластеров играет этот показатель.

Практический пример. Для иллюстрации практического применения описанного метода были взяты данные медицинского исследования. Были обследованы 28 пациентов с четырьмя числовыми (кол-во тромбоцитов, тромбиновое время (ТВ), лейкоциты, активированное частичное тромбопластиновое время (АЧТВ)) и двумя качественными показателями (аллели генов F5 Лейден и MTHFR) с 3 категориями каждый (нормозигота, гетерозигота, гомозигота). Все пациенты изначально были разделены медицинским экспертом в области заболевания тромбозами на 5 групп в соответствии со степенью тяжести их заболевания так, что увеличение номера группы соответствовало более тяжелой форме заболевания.

После оцифровки двух качественных показателей методом анализа соответствий каждая из категорий признаков получила свою векторную метку. Путем выбора в качестве числовой метки первой координаты как наиболее информативной, получили:

Таблица 1

Числовые метки

признак / **категория**	**F5 Лейден**	**MTHFR**
нормозигота	-8	-22
гетерозигота	-42	43
гомозигота	117	160

Таким образом, в нашем распоряжении оказалось 6 числовых показателей $X_1,...,X_6$. Вычисляя коэффициенты корреляции между всеми имеющимися показателями и кластерной переменной f поочередно, получили:

Таблица 2

Коэффициенты корреляции

X	***ρ***	**степень значимости**
кол-во тромбоцитов	-0.49	1
ТВ	0.14	5
АЧТВ	-0.32	3
лейкоциты	-0.14	6
F5 Лейден	0.21	4
MTHFR	-0.33	2

Из табл. 2 видим, что наиболее значимым показателем здесь является количество тромбоцитов, а наименее значимым — лейкоциты.

Верификация результатов исследования. Для верификации результата исследования применим метод оценки степени влияния числового показателя на вид кластерной структуры, предложенный в [3]. Метод оценки степени значимости влияния показателей на кластерную структуру (и, как следствие, решения post-hoc задачи) является прямым и не вызывает сомнений в своей объективности. Показатели ранжируются там по величине коэффициента кластерных различий разбиений, получаемых по полному набору показателей и после удаления из этого набора изучаемого показателя. Изучаемый показатель оказывается тем важнее, чем больше вычисленный коэффициент отличается от единицы.

Следуя [3], вычислим коэффициент кластерных различий между первоначальным разбиением и разбиениями, полученными при удалении признаков. Удаляя наиболее значимый показатель (количество тромбоцитов) и наименее значимый из всех показатель (лейкоциты), получили коэффициенты кластерных различий с первоначальным разбиением $k_1 = 0.778$ и $k_2 = 1$ соответственно. Т.к. $k_2 = 1$, то разбиения полностью идентичны, следовательно, с точки зрения метода [3] лейкоциты вовсе не влияют на кластерную структуру первоначального множества, что подтверждает низкую значимость этого показателя. А т.к. $k_1 = 0.778$, то соответствующие разбиения все-таки имеют высокую степень схожести, и следовательно, количество тромбоцитов является наиболее значимым признаком из всех рассматриваемых, хотя и не имеет сильного влияния на кластерную структуру изучаемого множества [3]. Итак, заключения двух методов сравнения важности изучаемых показателей в основном совпадают.

Наиболее очевидный вывод, следующий из произведенного анализа, состоит в том, что, вероятнее всего, в данном случае для точной дифференциальной диагностики степени тяжести тромбозов следует отказаться от рассмотрения данных признаков и искать новые, связанные с кластерной структурой более тесным образом. Но медицинская наука уверена в том, что изучение данных признаков позволяет уверенно определить степень тяжести заболевания, а применение коэффициента корреляции, как показано, неадекватно описывает степень влияния показателей на кластерную структуру. Это позволяет предположить, что такое влияние существенно нелинейно.

Алгоритм моделирования. Попробуем найти выход из такой ситуации следующим образом. Сохраняя установленный экспертом порядок следования кластеров, откажемся от равномерной шкалы их меток. В рассматриваемом примере метки кластеров были взяты равными числам от 1 до 5 соответственно. Шаг ранга в этом случае был постоянным и равен единице. Попробуем менять не шаг ранга, а регулярно сами метки. В качестве метки для j-го кластера будем использовать значение $f(j)$, j=1,...m. Назовем функцию $f(j)$ функцией перехода. Если при выборе какой-то конкретной функции перехода f модуль коэффициента корреляции показателя X_i окажется статистически значимым, это укажет на линейный характер влияния $f^{-1}(X_i)$ на номер кластера.

Итак, пусть нам удалось найти строго монотонно возрастающую функцию с наибольшим по модулю коэффициентом корреляции $\rho = \rho(X,f)$ между показателем X и кластерной переменной, на j-м кластере равной значению $f(j)$. Тогда

$$f(j\_A) = \rho \frac{S_f}{S_X}\left(X_A - \overline{X}\right) + \overline{f},$$

где j_A — номер кластера, к которому относится объект A, X_A — значение показателя X на этом объекте. Поэтому для нахождения по значению X_A номера того кластера, к которому относится объект A, следует вычислить величину

$$Z_X = f^{-1}\left(\rho \frac{S_f}{S_X}\left(X_A - \overline{X}\right) + \overline{f}\right).$$

Естественно, это можно сделать и для каждого из $p+q$ показателей. Полная прогностическая функция строится суммированием отдельных таких Z_X. Для более высокой точности можно учесть абсолютные величины коэффициентов корреляций, например, строя прогностическую функцию по формуле

$$\delta = \sum_{k=1}^{p+q} |\rho_k| Z_{X_k},$$

где ρ_k — соответствующий максимальный по модулю коэффициент корреляции для k-го показателя. Таким образом, мы получили некоторое число, с помощью которого после его нормировки (для попадания в интервал между $f(1)$ и $f(m)$) и округления до ближайшего целого, можно интерпретировать результат, т.е. мы получим номер кластера, к которому относится рассматриваемый объект.

Выводы. Используя априорную информацию о порядке следования кластеров, был предложен метод ранжирования определяющих показателей объектов. Поскольку было заранее известно, что каждый из исследуемых показателей существенно влияет на формирование кластеров, то предполагается существование «достойной» дискриминационной функции перехода, правильно разделяющей объекты по имеющимся кластерам. В силу проведенного исследования, показавшего несущественные влияния некоторых показателей посредством коэффициента корреляции, становится ясно, что эти показатели осуществляют свое влияние существенно нелинейным образом. Нами разработан алгоритм, позволяющий для каждого показателя определить вид функции перехода, посредством которой его влияние выделяется наиболее «правильным» образом.

СПИСОК ЛИТЕРАТУРЫ

1. Дронов С.В., Герасимова А.С. К проблеме оцифровки кластерной переменной / Тр. всеросс. молодежной школы-семинара «Анализ, геометрия и топология». Барнаул, 2013. С. 54-58.

2. Герасимова А.С. Кластеризация объектов с качественными категорированными признаками // Современная школа России. Вопросы модернизации: М-лы III Междунар. науч.-практич. конф. Москва: Открытый мир, 2013. С. 6-9.

3. Герасимова А.С. Кластеризация объектов с качественными признаками и ее использование для оценки силы их связи // Известия Алтайского государственного университета. Вып. 1/2(77). 2013. С. 66-69.

4. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.

5. Дронов С.В. Многомерный статистический анализ: Учебное пособие. Барнаул: Изд-во Алт. гос. ун-та, 2006. 221 с.

REFERENCES

1. Dronov, S.V., Gerasimova, A.S. On the problem of digitization of a cluster variable. *Tr. vseross. molodezhnoi shkoly-seminara «Analiz, geometriia i topologiia»* [Proceedings of Analysis, Geometry and Topology National Youth Workshop]. Barnaul, 2013. Pp. 54-58. (in Russian).

2. Gerasimova, A.S. Clusterization of the objects with non-numerical features. *Sovremennaia shkola Rossii. Voprosy modernizatsii: M-ly III Mezhdunar. nauch.-praktich. konf.* [Modern school of Russia. Modernization issues]. Moscow, 2013. Pp. 6-9. (in Russian).

3. Gerasimova, A.S. Clusterization of the objects with non-numerical features and its use for the estimation of their connection strength. *Izvestiia Altaiskogo gosudarstvennogo universiteta — Proceedings of Altai State University*. 2013. Issue 1/2 (77). Pp. 66-69. (in Russian).

4. Aivazian, S.A., Bukhshtaber, V.M., Eniukov, I.S., Meshalkin, L.D. *Prikladnaia statistika: Klassifikatsiia i snizhenie razmernosti* [Applied statistics: classification and reduction of dimension]. Moscow, 1989. 607 p. (in Russian).

5. Dronov, S.V. *Mnogomernyi statisticheskii analiz: Uchebnoe posobie* [Multidimensional statistical analysis: textbook]. Barnaul, 2006. 221 p. (in Russian).

Авторы публикации

Дронов Сергей Вадимович — доцент кафедры математического анализа факультета математики и информационных технологий Алтайского государственного университета, кандидат физико-математических наук (Барнаул)

Сазонова Антонина Станиславовна — аспирантка кафедры математического анализа факультета математики и информационных технологий Алтайского государственного университета (Барнаул)

Authors of the publication

Sergey V. Dronov — Cand. Sci. (Phys.-Math.), Associate Professor, Department of Mathematics and Information Technologies, Altai State University (Barnaul)

Antonina S. Sazonova — Post-graduate student, Department of Mathematics and Information Technologies, Altai State University (Barnaul)